

# CSP554—Big Data Technologies

## Initial Project Proposal

- **Project Topic** – To Build a Data Pipeline using Big Data SQL Technologies.

- **Area of Project** – Formula 1 / Auto Racing Sport.

- **Dataset** – Formula 1 Dataset –

The dataset is divided into Four csv files named circuits, constructors, drivers and driverGrid from year 1950 to 2022. Few of the important features are driverRef, points, position, wins, constructor\_ref, start\_position, constructor\_positions, etc.

<https://www.kaggle.com/datasets/harrybassi13/formula-1>

- **Problem Statement** –

To build a data processing pipeline using Big Data technology to insert, store, clean and use various data techniques to find out interesting insights / analytics from the Formula 1 Dataset.

- **Proposed Approach for the Solution** –

Database – Google Firestore database

Data Clean and Transformation – Google Big Query (Views)

Data Visualization – Data Studio by Google

Data Mining / Analysis – SQL

- **References** –

- <https://www.kaggle.com/datasets/harrybassi13/formula-1>
- <https://cloud.google.com/bigquery/docs>
- <https://developers.google.com/looker-studio>